Estimating *E. coli* concentrations in irrigation pond waters with machine learning algorithms

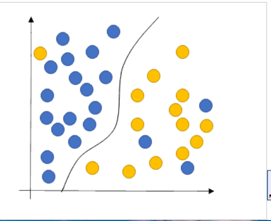
Matthew Stocker

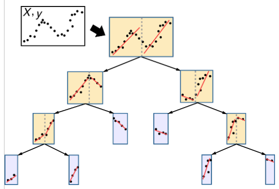
Yakov Pachepsky

Robert Hill

USDA ARS

DEPARTMENT OF ENVIRONMENTAL SCIENCE & TECHNOLOGY
www.enst.umd.edu





Farm Ponds and Food Safety

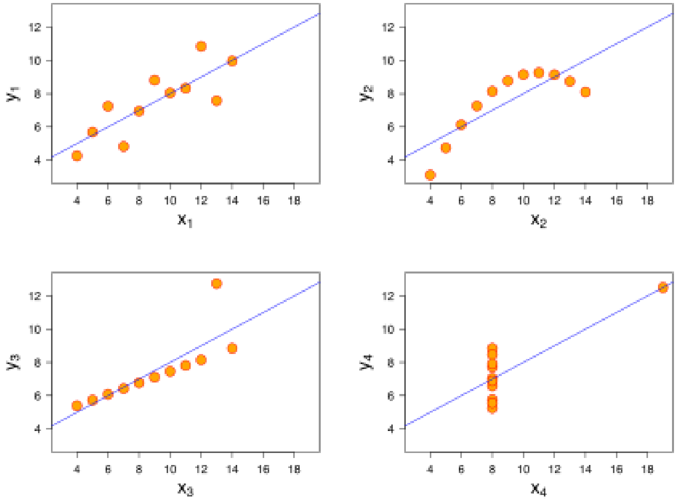
- Irrigation ponds constructed on farms are a convenient way to provide water for crop irrigation
- In 1980, the NRCS reported that there were over 2.1 million farm ponds constructed on private lands to provide irrigation water
- The quality of the waters withdrawn for irrigation are extremely important to food safety and public health
- During irrigation, microbial pathogens and chemicals contained in pond waters may be transferred to crops and their soils
- These substances can persist in the environment on plant surfaces or become internalized and persist during harvest and transport
- Farm pond water has been linked to foodborne outbreaks in the United States

Microbial water quality modeling

Determination of *E. coli* levels currently tedious and not timely

Researchers try to develop correlative relationships with water quality parameters and develops regression models

Strong associations are rarely reported between *E. coli* or pathogen concentrations and predictors



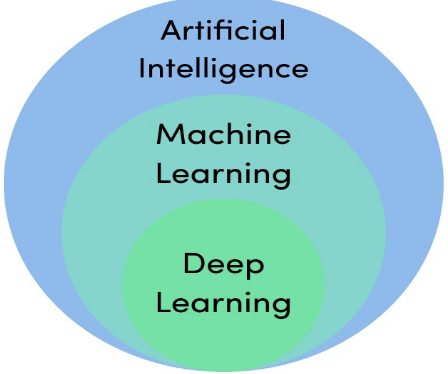
Machine Learning

Machine learning is a branch of artificial intelligence

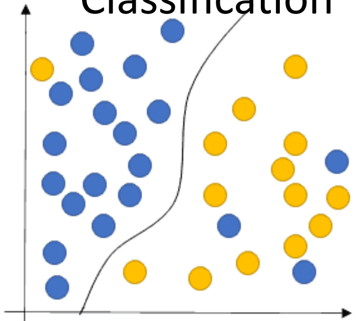
Powerful algorithms capable of “learning” from large and complex datasets

No assumptions of the nature of the dataset

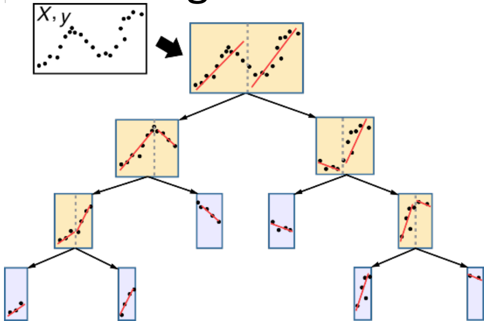
Shows promise for improving our predictive capabilities for water quality monitoring



Classification



Regression



Objectives:

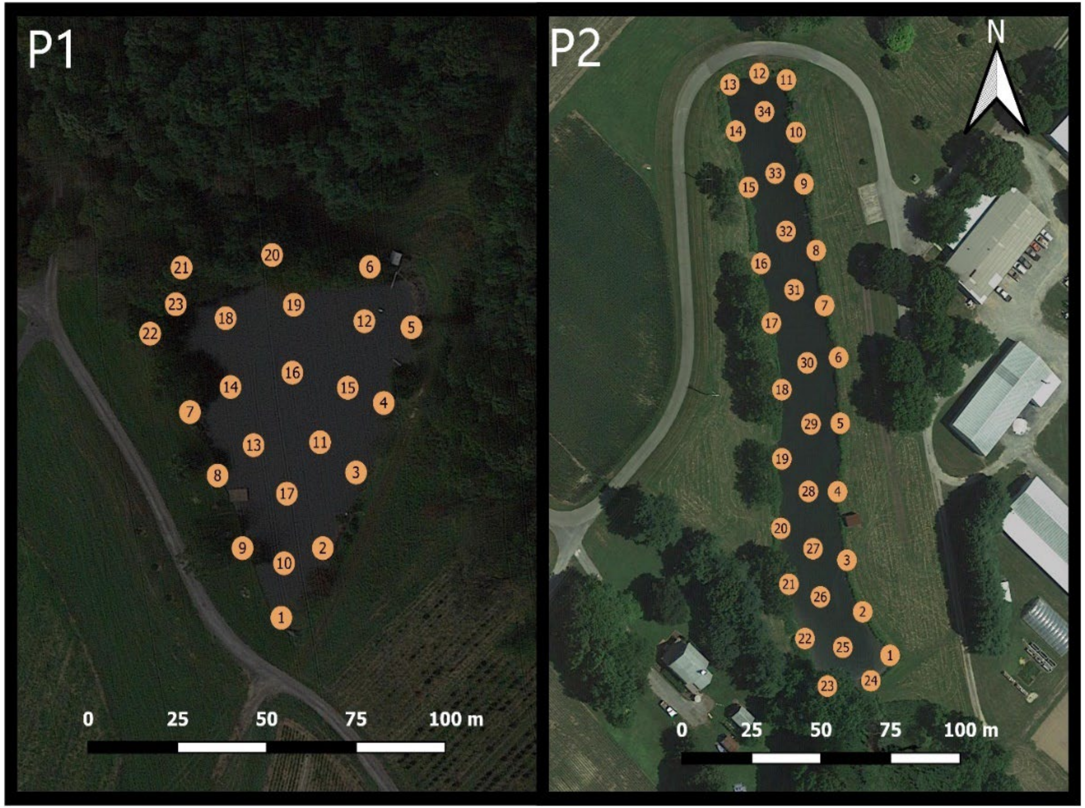
- Evaluate several machine learning algorithms for the prediction of *E. coli* concentrations using readily sensed/measured water quality data

- Test several sets of water quality variable sets from simple to complex to see if performance is increased with more predictors

- Determine the most influential predictors of the *E. coli* concentrations

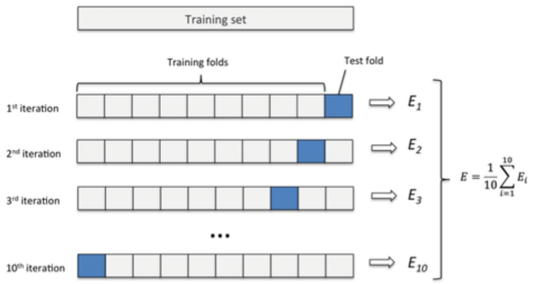
Locations and Sampling

- 2016 to 2018 dataset used
- Biweekly sampling during growing season
- Sampling in the morning hours
- Sampling avoided rainfall
- 23 and 34 locations in P1 and P2, respectively
- *E. coli* and water quality measurements recorded at each location

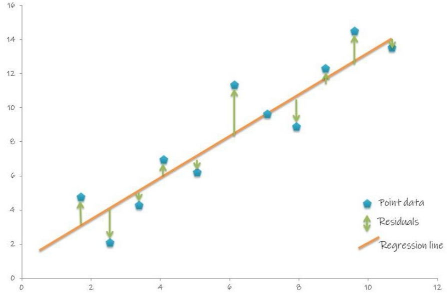


Scenario testing and error assessment

- Three sets of predictors
  - **Set A**: 5 parameter: DO, pH, NTU, °C, SPC (simple)
  - **Set AB**: 8 parameter: **Set A** + CHL, PC, FDOM (intermediate)
  - **Set ABC**: 12 parameters : **Set AB** + OP, NH_4+, TN, TC (complex)



- Random Forest (RF), k-Nearest Neighbors (*k*NN), Gradient boosting Machines (GBM), Support Vector Machines (SVM), Multiple Linear Regression (MLR)



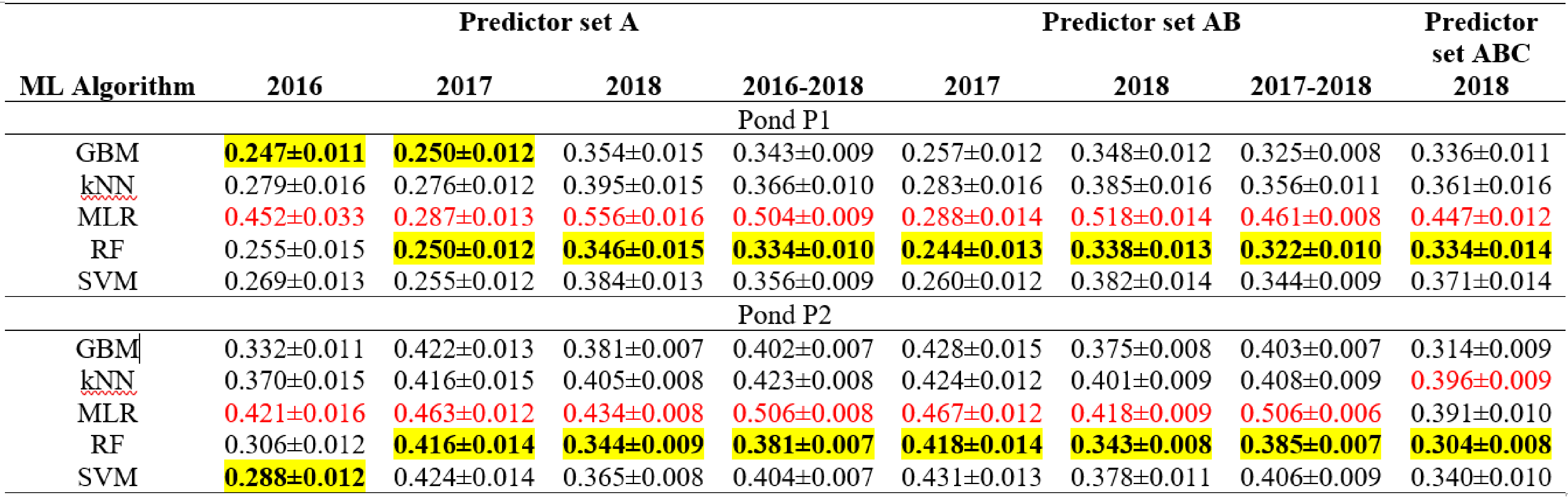
- Training and testing sets
  - Dataset comprised of over 900 samples (~ 8000 unique measurements)
  - 10-fold cross validation with 5 repeats was used

- Several metrics used for comparison, today will focus on values of Root-Mean-Square Error (RMSE)

- Minimal pre-processing

- Hyperparameter tuning to optimize accuracy

RMSE values



	Predictor set A				Predictor set AB			Predictor set ABC
ML Algorithm	2016	2017	2018	2016-2018	2017	2018	2017-2018	2018
Pond P1								
GBM	**0.247±0.011**	**0.250±0.012**	0.354±0.015	0.343±0.009	0.257±0.012	0.348±0.012	0.325±0.008	0.336±0.011
kNN	0.279±0.016	0.276±0.012	0.395±0.015	0.366±0.010	0.283±0.016	0.385±0.016	0.356±0.011	0.361±0.016
MLR	0.452±0.033	0.287±0.013	0.556±0.016	0.504±0.009	0.288±0.014	0.518±0.014	0.461±0.008	0.447±0.012
RF	0.255±0.015	**0.250±0.012**	**0.346±0.015**	**0.334±0.010**	**0.244±0.013**	**0.338±0.013**	**0.322±0.010**	**0.334±0.014**
SVM	0.269±0.013	0.255±0.012	0.384±0.013	0.356±0.009	0.260±0.012	0.382±0.014	0.344±0.009	0.371±0.014
Pond P2								
GBM	0.332±0.011	0.422±0.013	0.381±0.007	0.402±0.007	0.428±0.015	0.375±0.008	0.403±0.007	0.314±0.009
kNN	0.370±0.015	0.416±0.015	0.405±0.008	0.423±0.008	0.424±0.012	0.401±0.009	0.408±0.009	0.396±0.009
MLR	0.421±0.016	0.463±0.012	0.434±0.008	0.506±0.008	0.467±0.012	0.418±0.009	0.506±0.006	0.391±0.010
RF	0.306±0.012	**0.416±0.014**	**0.344±0.009**	**0.381±0.007**	**0.418±0.014**	**0.343±0.008**	**0.385±0.007**	**0.304±0.008**
SVM	**0.288±0.012**	0.424±0.014	0.365±0.008	0.404±0.007	0.431±0.013	0.378±0.011	0.406±0.009	0.340±0.010

Set A: 5 parameter: DO, pH, NTU, °C, SPC (minimum)
Set AB: 8 parameter: DO, pH, NTU, ° C, SPC + CHL, PC, FDOM (intermediate)
Set ABC: 12 parameters : DO, pH, NTU, ° C, SPC, CHL, PC, FDOM + OP, NH3, TN, TC (maximum)

GBM = gradient boosting machines, kNN = k-nearest neighbor, MLR = multiple linear regression,
RF = random forest, SVM = support vector machines

Recursive feature elimination – Top 5 predictors for each variable set

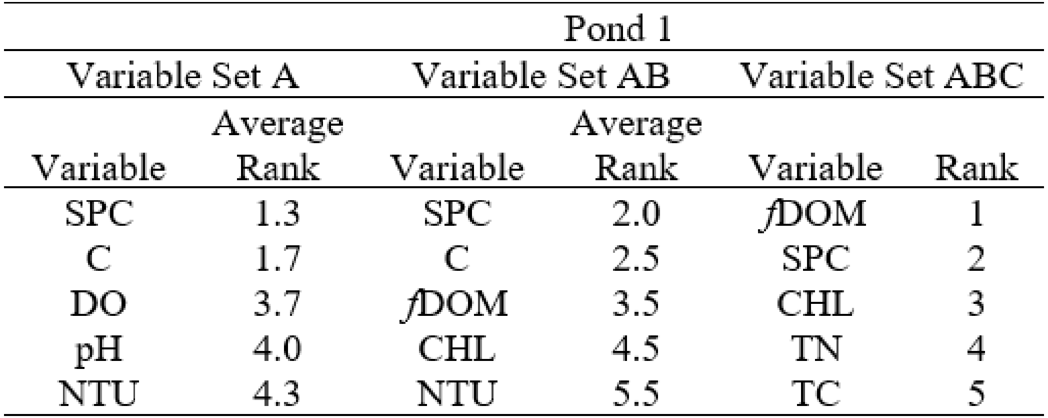
- SPC important predictor in all scenarios

- °C second most important in A and AB

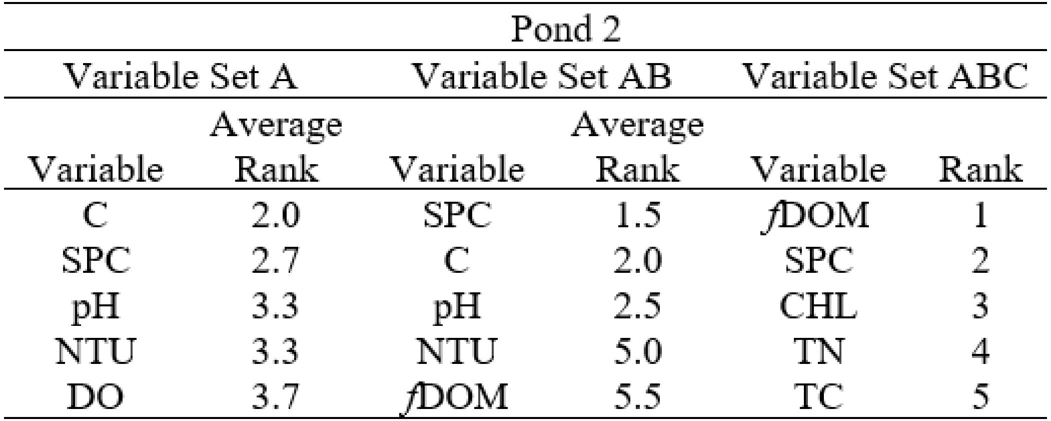
- ABC has FDOM at top and newly added total nutrient data in top 5 of 12

- P2 also has SPC and C as most influential in the A and AB scenarios

- Exact same importance ranking in the ABC scenario as P1
  - Important to note there is no direction of the relationship presented
    - Nonlinear relationship containing threshold values of certain parameters

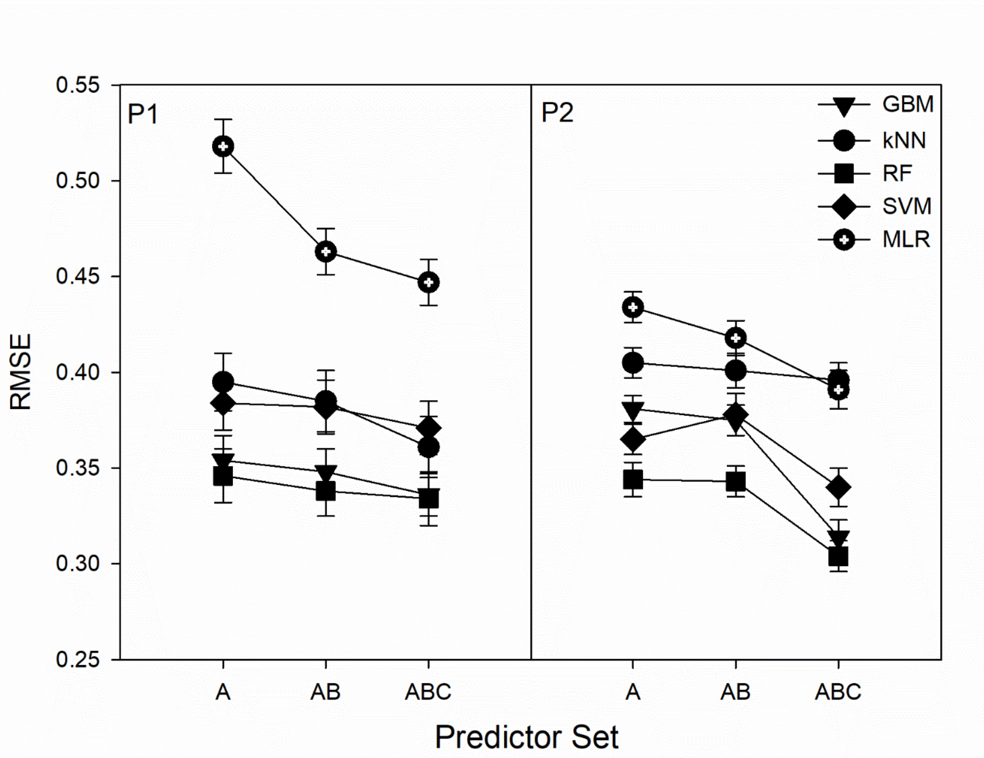


Pond 1					
Variable Set A		Variable Set AB		Variable Set ABC	
	Average		Average		
Variable	Rank	Variable	Rank	Variable	Rank
SPC	1.3	SPC	2.0	*f*DOM	1
C	1.7	C	2.5	SPC	2
DO	3.7	*f*DOM	3.5	CHL	3
pH	4.0	CHL	4.5	TN	4
NTU	4.3	NTU	5.5	TC	5



Pond 2					
Variable Set A		Variable Set AB		Variable Set ABC	
	Average		Average		
Variable	Rank	Variable	Rank	Variable	Rank
C	2.0	SPC	1.5	*f*DOM	1
SPC	2.7	C	2.0	SPC	2
pH	3.3	pH	2.5	CHL	3
NTU	3.3	NTU	5.0	TN	4
DO	3.7	*f*DOM	5.5	TC	5

RMSE dependence on predictor sets in the 2018 datasets



Set A: 5 parameter: DO, pH, NTU, °C, SPC

Set AB: 8 parameter: DO, pH, NTU, ° C, SPC + CHL, PC, FDOM

Set ABC: 12 parameters : DO, pH, NTU, °C, SPC, CHL, PC, FDOM + OP, NH4, TN, TC

Conclusions and Recommendations

- Random forest by in large performed the best across the datasets for each pond and scenarios tested
  - Best model is dataset-dependent, as somewhat evidenced in this work

- Model performance was rarely significantly different, though there was consistency in best model

- The 2018 A/AB/ABC datasets showed that the simpler suite of variables provided results comparable to when the 12 parameter sets were used as inputs
  - Implication that a simple and cheaper set of sensors may be adequate for making predictions

- Models created using machine learning algorithms show promise for conducting quick or real-time microbial water quality assessments based on continuously sensed water quality data
  - Provide accurate predictions on highly complex and non-linear datasets which traditional statistical models may struggle with

Acknowledgements

- Members of the Environmental Microbial Food Safety Laboratory and ARS
  - Jaclyn Smith
  - Billie Morgan
  - Laura Del Collo
  - Jakeitha Sonnier
  - Dr. Jo Ann van Kessel
  - Dr. Monica Santin-Duran
  - George Meyers
- Farm Staff and Research Directors
  - Mike Newell
  - Dr. Kenneth Staver
  - Dr. Kate Everts
  - Private property owners
- Interns
  - Mauricio Peña
  - Cesar Hernandez
  - Ashley Vazquez
  - Nhu Lee
  - Nailah Washington
  - Lauren Wyatt-Brown

DEPARTMENT OF ENVIRONMENTAL SCIENCE & TECHNOLOGY
www.enst.umd.edu

USDA ARS

ORISE

Stocker, M. D., Pachepsky, Y. A., & Hill, R. L. (2022). Prediction of *E. coli* Concentrations in Agricultural Pond Waters: Application and Comparison of Machine Learning Algorithms. *Frontiers in artificial intelligence*, *4*, 768650.

Questions?